SIAM J. MATRIX ANAL. APPL.
Vol. 30, No. 4, pp. 1773–1788

ON THE CONVERGENCE OF THE SELF-CONSISTENT FIELD ITERATION FOR A CLASS OF NONLINEAR EIGENVALUE PROBLEMS[*]

CHAO YANG[†], WEIGUO GAO[‡], AND JUAN C. MEZA[†]

Abstract. We investigate the convergence of the self-consistent field (SCF) iteration used to solve a class of nonlinear eigenvalue problems. We show that for the class of problems considered, the SCF iteration produces a sequence of approximate solutions that contain two convergent subsequences. These subsequences may converge to two different limit points, neither of which is the solution to the nonlinear eigenvalue problem. We identify the condition under which the SCF iteration becomes a contractive fixed point iteration that guarantees its convergence. This condition is characterized by an upper bound placed on a parameter that weighs the contribution from the nonlinear component of the eigenvalue problem. We derive such a bound for the general case as well as for a special case in which the dimension of the problem is 2.

Key words. nonlinear eigenvalue problem, self-consistent field iteration polynomial

AMS subject classifications. 15A18, 65F15, 47J10

DOI. 10.1137/080716293

1. Introduction.

We are concerned with the convergence of a numerical method for solving the following type of nonlinear eigenvalue problem:

$$H(X)X = X\Lambda_k, \tag{1}$$

where $X \in \mathbb{R}^{n\times k}$, $X^TX = I_k$, $H(X) \in \mathbb{R}^{n\times n}$ is a matrix that has a special structure to be defined below, and $\Lambda_k \in \mathbb{R}^{k\times k}$ is a diagonal matrix consisting of the k smallest eigenvalues of $H(X)$. This type of problem arises in electronic structure calculations [10, 6]. The nonlinearity simply refers to the dependency of the matrix H on the eigenvector X to be computed. This dependency is expressed through a vector $\rho(X)$ that corresponds to the *charge density* of electrons in an electronic structure calculation. This vector is defined as

$$\rho(X) \equiv \mathrm{diag}(XX^T), \tag{2}$$

where $\mathrm{diag}(A)$ denotes the vector containing the diagonal elements of the matrix A.

Given $\rho(X)$, the matrix $H(X)$ that we will consider in this paper is defined as

$$H(X) = L + \alpha \mathrm{Diag}(L^{-1}\rho(X)), \tag{3}$$

where L is a discrete Laplacian, $\mathrm{Diag}(x)$ (with an uppercase D) denotes a diagonal matrix with x on its diagonal, and α is some known constant. In electronic structure calculations, $H(X)$ is often referred to as a single-particle Hamiltonian.

[*] Received by the editors February 22, 2008; accepted for publication (in revised form) by M. E. Hochstenbach October 20, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/simax/30-4/71629.html

[†] Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 (CYang@lbl.gov, JCMeza@lbl.gov). The work of these authors was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract DE-AC03-76SF00098.

[‡] School of Mathematical Sciences, Fudan University, Shanghai, People's Republic of China (wggao@fudan.edu.cn). This author's research was supported by the Special Funds for Major State Basic Research Projects of China (grant 2005CB321700) and the 111 Project.

The solution of (1) is also a global minimizer of the constrained minimization problem

$$\begin{array}{ll} \min & E(X) \\ \text{s.t.} & X^T X = I_k, \end{array} \tag{4}$$

where the objective function $E(X)$ is defined by

$$E(X) = \frac{1}{2}\text{trace}(X^T L X) + \frac{\alpha}{4}\rho(X)^T L^{-1}\rho(X). \tag{5}$$

In fact, it is not difficult to show that (1) and the orthonormality constraint $X^T X = I_k$ form the first order necessary conditions for (4) [7].

The nonlinear eigenvalue problem defined by (1) and (3) is a simplification of the Hartree–Fock (HF) and Kohn–Sham (KS) equations in electronic structure calculations [10, 6]. In particular, it contains a parameterized Hartree term $\rho^T L^{-1}\rho$ that is present in both the HF and KS equations. But it does not contain the exchange term in the HF model [10] or the exchange-correlation term in the KS model [6]. Although our analysis is performed on this simplified model, the main results reveal some of the fundamental properties of this type of problem and how the behavior of the algorithm used to solve this type of problem changes with respect to the amount of nonlinearity measured by the parameter α in (3).

The numerical method we will analyze is called the *self-consistent field* (SCF) iteration. It is currently the most widely used algorithm for solving the HF and KS equations. In each SCF iteration, one computes approximations to a few of the smallest eigenvalues and the corresponding eigenvectors of a fixed Hamiltonian constructed from the previous approximation to X; the computed eigenvector approximations are used to update the Hamiltonian. When the difference between Hamiltonians constructed in two consecutive iterations is negligible, the SCF procedure is terminated, and the eigenvectors of the last Hamiltonian are said to be self-consistent.

It is well known that the simplest version of SCF iteration, which we will carefully describe in the next section, often fails to converge [5]. For certain types of Hamiltonians (e.g., HF and the one defined in (3)), the SCF iteration may eventually oscillate between two limit points, neither of which satisfies (1). The convergence failure of the SCF iteration is partially explained in [11] by viewing the SCF iteration as an indirect minimization procedure that seeks the minimum of (4) by minimizing a sequence of quadratic surrogates. Although the arguments and numerical examples presented in [11] demonstrated that $E(X)$ may not decrease monotonically in an SCF iteration, they do not reveal the asymptotic convergence behavior of the SCF iteration.

In this paper, we will take a closer look at the SCF iteration and analyze its convergence when used to solve (1). A brief overview of the algorithm is given in section 2 along with a simple example that illustrates the convergence failure of the SCF iteration for some choices of α used in (3). In section 3, we show that when the SCF iteration fails to converge, the approximate eigenvectors $X^{(i)}$ produced in the SCF iteration contain two subsequences that converge to two distinct limit points. Neither of these limit points is a solution to (1). Our proof of this result is similar to an earlier proof given by Cancès and Le Bris in [2]. We made a number of simplifications to make it easier to follow. However, the subsequence convergence result does not give the conditions under which the two subsequences are guaranteed to converge to the solution of (1). Such a condition is identified in section 4. We will show that for $n = 2$, the SCF iteration is guaranteed to converge when $\alpha < 3$. For the more general case,

```
SCF Iteration

Input:  A discrete Laplacian $L \in \mathbb{R}^{n\times n}$; an initial guess $X^{(0)}$ for the eigenvector $X \in \mathbb{R}^{n\times k}$;
Output: $X \in \mathbb{R}^{n\times k}$ such that $X^*X = I_k$ and $H(X)X = X\Lambda_k$, where $\Lambda_k$ contains the $k$ smallest eigenvalues of $H(X)$.

For $i = 1, 2, \ldots$ until convergence
   construct $H^{(i)} = H(X^{(i-1)})$ using (3);
   compute $X^{(i)}$ such that $H^{(i)}X^{(i)} = X^{(i)}\Lambda^{(i)}$, and $\Lambda^{(i)}$
      contains the $k$ smallest eigenvalues of $H^{(i)}$;
end for
```

FIG. 1. *The SCF iteration.*

our main result provides an upper bound for α that depends on the minimum gap between the kth and the $k+1$st eigenvalues of $H(X)$, the dimension of the problem, and the norm of L^{-1}.

Throughout this paper, we will use $\|\cdot\|_p$ to denote the p-norm [3] of either a vector or a matrix. The Frobenius norm of a matrix is denoted by $\|\cdot\|_F$.

2. The SCF iteration.

In this section, we describe the SCF iteration and show how it fails when it is applied to a 2×2 Hamiltonian (3) with a particular choice of α.

The basic idea of an SCF iteration is to reduce the nonlinear eigenvalue problem (1) to a sequence of linear eigenvalue problems that can be solved efficiently using existing tools. Figure 1 shows the main steps of this procedure. The convergence of the iteration can be monitored by computing the difference between charge densities $\rho(X)$ obtained in two consecutive iterations. The following example shows that the simplest version of the SCF iteration fails to converge. In this example, we set

$$L = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \tag{6}$$

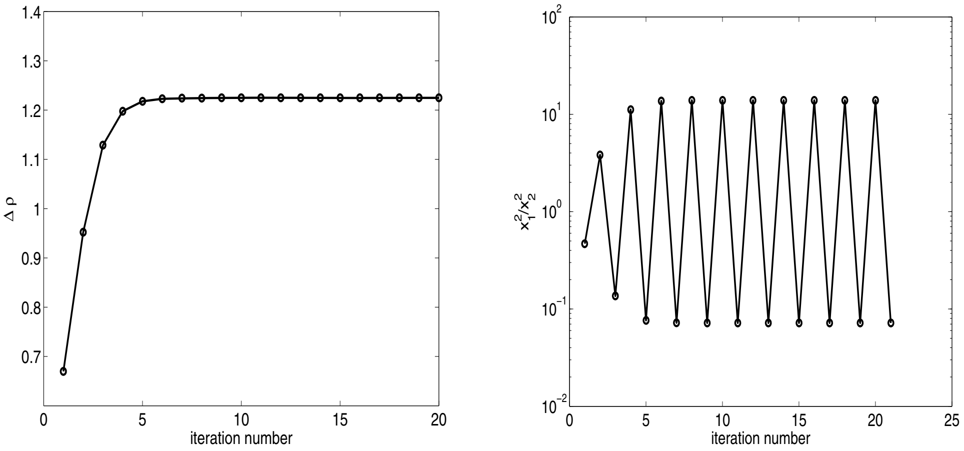
$\alpha = 12$, and $k = 1$. As a result, $X = (x_1 \;\; x_2)^T$ with $x_1, x_2 \in \mathbb{R}$ such that $x_1^2 + x_2^2 = 1$, and $\rho(X) = (x_1^2 \;\; x_2^2)^T$.

Due to the convexity and symmetry of $E(x)$ (i.e., interchanging x_1 and x_2 does not change the problem), the solution to the minimization problem (4), and hence the nonlinear eigenvalue problem (1), must satisfy $x_1 = x_2 = \sqrt{2}/2$ or $x_1 = x_2 = -\sqrt{2}/2$.

However, when the initial guess of the desired eigenvector is chosen to be, for example,

$$X^{(0)} = \begin{pmatrix} 0.1389 \\ 0.2028 \end{pmatrix}, \tag{7}$$

the difference between the charge densities computed in two consecutive SCF iterations does not converge to zero, as we can clearly see in Figure 2(a). Furthermore, Figure 2(b) shows that the ratio between two components of $\rho(X^{(i)})$ does not converge to 1.



(a) The change in charge density $\Delta\rho^{(i)} \equiv \|\rho(X^{(i+1)}) - \rho(X^{(i)})\|_2$ fails to converge to zero.

(b) The ratio between x_1^2 and x_2^2 oscillates around one, but does not converge to one.

FIG. 2. *The SCF iteration fails to converge when* $\alpha = 12$ *in* (3).

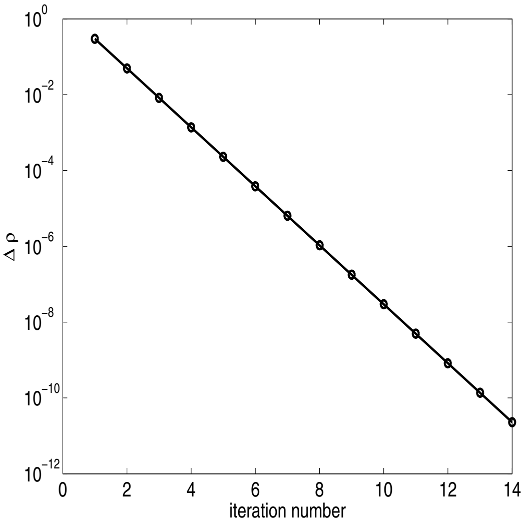
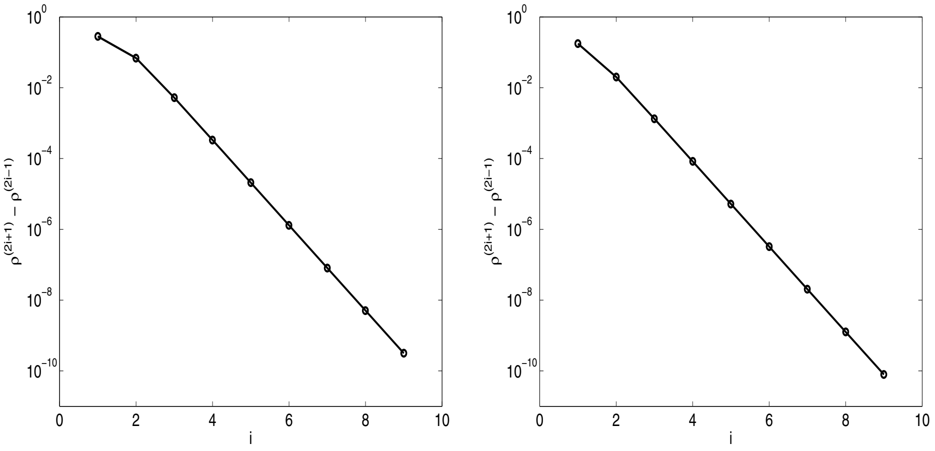


FIG. 3. *When* $\alpha = 1.0$, $\Delta\rho^{(i)}$ *converges rapidly to* 0.

If we reduce α to 1, then SCF converges from any starting guess. Figure 3 shows that the difference between charge densities computed in two consecutive SCF iterations decreases rapidly towards zero in this case when (7) is used as the starting guess. In section 4, we will show that for this 2×2 example, the convergence of SCF can be guaranteed if $\alpha < 3$.

3. Subsequence convergence in the SCF iteration. When the SCF iteration fails to converge to the solution of (1), it produces a sequence of approximations



(a) Odd iterations

(b) Even iterations

FIG. 4. *When $\alpha = 12$, the charge density converges to two different limit points in odd and even SCF iterations.*

$\{X^{(i)}\}$ that do not become self-consistent as i increases. We have already seen this phenomenon in Figure 2(a), where we plotted the norm of the change in $\rho(X^{(i)})$ between two consecutive SCF iterations. In this case, it is clear that $\|\Delta\rho(X^{(i)})\|_2$ does not converge to zero as i increases.

However, if we examine the subsequences $\{X^{(2i-1)}\}$ and $\{X^{(2i)}\}$ $(i = 1, 2, \ldots)$ produced in the SCF iteration, we will see that they both converge to subspaces that become self-consistent in every other iteration. Figure 4 shows that both

$$\Delta\rho^{(i)}_{\text{odd}} \equiv \|\rho(X^{(2i+1)}) - \rho(X^{(2i-1)})\|_2 \quad \text{and} \quad \Delta\rho^{(i)}_{\text{even}} \equiv \|\rho(X^{(2i+2)}) - \rho(X^{(2i)})\|_2$$

converge to zero as i increases, although neither $X^{(2i+1)}$ nor $X^{(2i+2)}$ becomes a minimizer of $E(X)$, as we can clearly see in Figure 2(b).

In [1] and [2] it was shown that such a phenomenon occurs in a more general setting; i.e., when SCF fails to converge to the solution of the HF equation, the odd and even subsequences of the approximations converge to two distinct limit points. This analysis, which we will reproduce here with some modification, is based on examining the convergence of the density matrix $D(X) = XX^T$. It relies on the assumption that there exists a gap δ between the kth and $k+1$st eigenvalues of $H(X)$ for all valid X, an assumption that is referred to in [1] as the *uniformly well posed* (UWP) property. The major result of [1] asserts that

$$\sum_{i=\ell}^{\infty} \|D(X^{(i+2)}) - D(X^{(i)})\|_F^2 < \infty$$

for any finite $\ell \geq 0$. Therefore, $\|D(X^{(i+2)}) - D(X^{(i)})\|_F$ must converge to zero as i increases.

In the analysis we present next, the subsequence convergence of the SCF iteration is measured by the distance between two subspaces spanned by columns of $X \in \mathbb{R}^{n\times k}$ and $Y \in \mathbb{R}^{n\times k}$. We will use the standard distance measure defined in

[3, Theorem 2.6.1, p. 76]; i.e., if $X^TX = Y^TY = I_k$,

$$\operatorname{dist}(X,Y) \equiv \|Z^TY\|_2,$$

where $Z \in \mathbb{R}^{n\times(n-k)}$ is the orthogonal complement to X and $Z^TZ = I_{n-k}$.

The following lemma, which is a block version of Lemma 11-9-8 in [8], shows that $\operatorname{dist}(X,Y)$ can, in general, be bounded in terms of $\operatorname{trace}(Y^THY) - \operatorname{trace}(X^THX)$ and the gap between the kth and $k+1$st eigenvalues of H if columns of X consist of eigenvectors associated with the k smallest eigenvalues of H and $Y \in \mathbb{R}^{n\times k}$ satisfies $Y^TY = I_k$.

LEMMA 1. *Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be eigenvalues of a symmetric matrix $H \in \mathbb{R}^{n\times n}$, and let columns of X be eigenvectors associated with $\lambda_1, \lambda_2, \ldots, \lambda_k$. If $\lambda_{k+1} = \lambda_k + \delta$ for some $\delta > 0$, then*

$$\operatorname{dist}^2(X,Y) \le \frac{\operatorname{trace}(Y^THY) - \operatorname{trace}(X^THX)}{\delta} \tag{8}$$

for any $Y \in \mathbb{R}^{n\times k}$ such that $Y^TY = I_k$.

Proof. Let columns $Z \in \mathbb{R}^{n\times(n-k)}$ be eigenvectors associated with $\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_n$, and define $\Lambda_k = \operatorname{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ and $\Lambda_{n-k} = \operatorname{Diag}(\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_n)$. It follows from the spectral decomposition of H that

$$\operatorname{trace}(Y^THY) = \operatorname{trace}[(Y^TX)\Lambda_k(X^TY)] + \operatorname{trace}[(Y^TZ)\Lambda_{n-k}(Z^TY)].$$

Since $\lambda_{k+1} = \lambda_k + \delta$, we have $\lambda_i \ge \lambda_k + \delta$ for $i \ge k+1$. Thus,

$$\operatorname{trace}[(Y^TZ)\Lambda_{n-k}(Z^TY)] \ge (\lambda_k + \delta)\|Z^TY\|_F^2.$$

Consequently,

$$\operatorname{trace}(Y^THY) \ge \operatorname{trace}[(Y^TX)\Lambda_k(X^TY)] + \lambda_k\|Z^TY\|_F^2 + \delta\|Z^TY\|_F^2. \tag{9}$$

Because $W = (X, Z)$ defines an orthogonal transformation, we have

$$\|W^TY\|_F^2 = \|Y\|_F^2 = k.$$

Hence

$$\|Z^TY\|_F^2 = \|W^TY\|_F^2 - \|X^TY\|_F^2 = k - \|X^TY\|_F^2. \tag{10}$$

Substituting (10) into (9) and setting $S = X^TY$ yields

$$\begin{aligned}\operatorname{trace}(Y^THY) &\ge \operatorname{trace}(S\Lambda_kS^T) + \lambda_k(k - \|S\|_F^2) + \delta\|Z^TY\|_F^2\\ &= \lambda_k k + \operatorname{trace}[S(\Lambda_k - \lambda_k I)S^T] + \delta\|Z^TY\|_F^2\\ &= \operatorname{trace}(\Lambda_k) + \operatorname{trace}(\lambda_k I - \Lambda_k) + \operatorname{trace}[(\Lambda_k - \lambda_k I)SS^T] + \delta\|Z^TY\|_F^2\\ &= \operatorname{trace}(X^THX) + \operatorname{trace}[(\lambda_k I - \Lambda_k)(I - SS^T)] + \delta\|Z^TY\|_F^2.\end{aligned}$$

Because $X^TX = Y^TY = I_k$, the diagonal elements of SS^T are all less than or equal to one. Hence

$$\operatorname{trace}[(\lambda_k I - \Lambda_k)(I - SS^T)] \ge 0.$$

Therefore, we can now conclude that

$$\begin{aligned}\operatorname{trace}(Y^THY) &\geq \operatorname{trace}(X^THX) + \delta\|Z^TY\|_F^2\\ &\geq \operatorname{trace}(X^THX) + \delta\|Z^TY\|_2^2\\ &= \operatorname{trace}(X^THX) + \delta\operatorname{dist}^2(X,Y).\end{aligned}$$

Rearranging terms in the above inequality yields (8). □

Our analysis of the subsequence convergence will make use of the auxiliary function

$$\hat{E}(X,Y) = \operatorname{trace}(X^TLX) + \operatorname{trace}(Y^TLY) + \alpha\rho(X)^TL^{-1}\rho(Y). \tag{11}$$

This function is similar to the one used in [1], which is defined in terms of density matrices $D(X)$ and $D(Y)$.

It is easy to verify that

$$\rho(X)^TL^{-1}\rho(Y) = \operatorname{trace}(X^T\operatorname{Diag}[L^{-1}\rho(Y)]X) = \operatorname{trace}(Y^T\operatorname{Diag}[L^{-1}\rho(X)]Y).$$

Thus, $\hat{E}(X,Y)$ is clearly symmetric, i.e., $\hat{E}(X,Y) = \hat{E}(Y,X)$, and it can be expressed alternatively as

$$\begin{aligned}\hat{E}(X,Y) &= \operatorname{trace}(X^TH(Y)X) + \operatorname{trace}(Y^TLY)\\ &= \operatorname{trace}(Y^TH(X)Y) + \operatorname{trace}(X^TLX). \qquad (12)\end{aligned}$$

We are now ready to show the main result, which we state formally in the following theorem.

THEOREM 1. *Let $X^{(0)} \in \mathbb{R}^{n\times k}$ be the initial guess to the solution of the nonlinear eigenvalue problem* (1) *that satisfies ${X^{(0)}}^TX^{(0)} = I_k$. If columns of $X^{(i)} \in \mathbb{R}^{n\times k}$ contain eigenvectors associated with the smallest k eigenvalues of $H(X^{(i-1)})$, as we would obtain when applying the SCF iteration to* (1)*, and if the gap between the kth and the $k+1$st eigenvalues of $H(X^{(i)})$ is greater than or equal to $\delta > 0$ for all i, then*

$$\sum_{i=0}^{m}\operatorname{dist}^2(X^{(i+2)},X^{(i)}) \leq \frac{\hat{E}(X^{(0)},X^{(1)}) - \hat{E}(X^{(m+1)},X^{(m+2)})}{\delta}, \tag{13}$$

where $\hat{E}(\cdot,\cdot)$ is the auxiliary function defined in (11).

Proof. The proof we give here is similar to that presented in [2]. To simplify notation, we will denote $H(X^{(i+1)})$ by H. Because $X^{(i+2)}$ contains eigenvectors associated with the smallest k eigenvalues of H, it follows from Lemma 1 that

$$\operatorname{trace}({X^{(i+2)}}^THX^{(i+2)}) + \delta\operatorname{dist}^2(X^{(i+2)},X^{(i)}) \leq \operatorname{trace}({X^{(i)}}^THX^{(i)}).$$

Adding $\operatorname{trace}({X^{(i+1)}}^TLX^{(i+1)})$ to both sides of the inequality above and invoking (12) yields

$$\hat{E}(X^{(i+1)},X^{(i+2)}) + \delta\operatorname{dist}^2(X^{(i+2)},X^{(i)}) \leq \hat{E}(X^{(i)},X^{(i+1)}).$$

Rearranging terms in the above inequality yields

$$\operatorname{dist}^2(X^{(i+2)},X^{(i)}) \leq \frac{\hat{E}(X^{(i)},X^{(i+1)}) - \hat{E}(X^{(i+1)},X^{(i+2)})}{\delta}.$$

Summing over i yields the inequality (13). □

Because $\hat{E}(X^{(m+1)},X^{(m+2)})$ can be bounded by a constant for any m, and the left-hand side of (13) is an increasing series, $\operatorname{dist}(X^{(i+2)},X^{(i)})$ must converge to zero as $i \to \infty$.

4. The convergence of SCF.

Although the subsequence convergence analysis characterizes what would happen when the SCF iteration fails to converge, it does not give the conditions under which both the even and odd subsequences are guaranteed to converge to the solution of (1). On the other hand, the numerical examples presented in section 2 appear to indicate that the convergence of SCF for the 2×2 problem depends on the value of α, which weighs the contribution of the nonlinear term $\mathrm{Diag}(L^{-1}\rho(X))$ in the Hamiltonian (3). In this section, we will provide a formal proof that this is indeed true. We will prove that the SCF iteration is guaranteed to converge to the solution of (1) from any starting point when $\alpha < \alpha_{\max}$ for some upper bound $\alpha_{\max}$.

Before we state and derive a general bound for α, we will first examine the convergence of the 2×2 problem shown in section 2 because this problem is relatively easy to analyze and because we can obtain a much tighter upper bound on α in this special case.

In section 4.2, we will use a more sophisticated technique to derive an upper bound for α that is more general but somewhat pessimistic.

4.1. The 2×2 case.

Before we get to the main result, we will first show that the ratio between the two components of the charge density oscillates around 1 regardless of the choice of α. We will later show that the magnitude of the oscillation decreases to zero when α is sufficiently small.

LEMMA 2. *Let $y = (y_1 \;\; y_2)^T$ be the eigenvector associated with the smallest eigenvalue of $H(X)$ defined in* (3)*, where $X = (x_1 \;\; x_2)^T$ with $|x_1| > |x_2|$. If $\alpha > 0$ in* (3)*, then $|y_2| > |y_1|$.*

Proof. It is straightforward to write down the inverse of L defined in (6) and show that

$$L^{-1}\rho(X) = \frac{1}{3}\begin{pmatrix} 2x_1^2 + x_2^2 \\ x_1^2 + 2x_2^2 \end{pmatrix}.$$

Consequently, the two diagonal elements in the second term of $H(X)$ in (3) are simply

$$\beta_1 = \frac{\alpha}{3}(2x_1^2 + x_2^2) \quad \text{and} \quad \beta_2 = \frac{\alpha}{3}(x_1^2 + 2x_2^2). \tag{14}$$

Suppose λ is an eigenvalue of $H(X)$; then

$$\det\begin{pmatrix} 2+\beta_1-\lambda & -1 \\ -1 & 2+\beta_2-\lambda \end{pmatrix} = (2+\beta_1-\lambda)(2+\beta_2-\lambda) - 1 = 0. \tag{15}$$

If we let $\phi(\lambda) = (2+\beta_1-\lambda)(2+\beta_2-\lambda)$, then eigenvalues of H are solutions to the equation $\phi(\lambda) = 1$.

It is easy to see from (14) that

$$\beta_1 - \beta_2 = \frac{\alpha}{3}(x_1^2 - x_2^2) > 0, \tag{16}$$

since $|x_1| > |x_2|$. Therefore, the two eigenvalues of $H(X)$, which are distinct roots of the quadratic equation $\phi(\lambda) = 1$, must satisfy

$$\lambda_1 < 2 + \beta_2 < 2 + \beta_1 < \lambda_2. \tag{17}$$

Let $y = (y_1 \;\; y_2)^T$ be the eigenvector associated with λ_1. It follows from $H(X)y = \lambda_1 y$ that

$$(2 + \beta_1 - \lambda_1)y_1 = y_2. \tag{18}$$

Because (17) implies $0 < 2 + \beta_2 - \lambda_1 < 2 + \beta_1 - \lambda_1$, it follows from (15) that

$$2 + \beta_1 - \lambda_1 > 1.$$

Consequently, we can deduce from (18) that $|y_2| > |y_1| > 0$. □

Lemma 3 confirms the observation we made in Figure 2(b), namely, that the ratio between the first and second components of ρ oscillates around 1 in the SCF iteration. The convergence of x_1 and x_2 to the optimal solution can be easily proved if we can show that

$$\frac{|y_2|}{|y_1|} < \frac{|x_1|}{|x_2|} \quad \text{when} \quad |x_1| > |x_2|, \tag{19}$$

or

$$\frac{|y_1|}{|y_2|} < \frac{|x_2|}{|x_1|} \quad \text{when} \quad |x_2| > |x_1|. \tag{20}$$

Without loss of generality, we will establish the condition under which (19) holds. However, before we do that, let us first express y_2/y_1 as a function of $\beta_1 - \beta_2$.

LEMMA 3. *If β_1 and β_2 are defined by* (14)*, then*

$$\frac{y_2}{y_1} = \frac{(\beta_1 - \beta_2) + \sqrt{(\beta_1 - \beta_2)^2 + 4}}{2}, \tag{21}$$

where $y = (y_1, y_2)^T$ is the eigenvector associated with the smallest eigenvalue of $H(X)$.

Proof. Let $\delta = y_2/y_1 = 2 + \beta_1 - \lambda_1$. It is easy to show that

$$2 + \beta_2 - \lambda_1 = \delta - (\beta_1 - \beta_2).$$

Hence, it follows from (15) that

$$\delta^2 - (\beta_1 - \beta_2)\delta - 1 = 0. \tag{22}$$

Solving (22) for δ and taking the positive root yields (21). □

Note that if $x_1 = x_2 = \sqrt{2}/2$, then $\beta_1 - \beta_2 = 0$. In this case, it follows from (21) that $y_2/y_1 = 1$, which matches our intuitive expectation that the SCF iteration should converge right away when the initial guess is the solution to (1).

The following theorem establishes the condition that guarantees the monotonic convergence of the SCF iteration when the initial guess is not the solution to (1).

THEOREM 2. *Let $X = (x_1 \;\; x_2)^T$ be an initial guess of the solution to* (1)*, where $H(X)$ is defined by* (3)*, and let $(y_1 \;\; y_2)^T$ be the eigenvector associated with the smallest eigenvalue of $H(X)$. If $|x_1| > |x_2|$, then*

$$\left|\frac{y_2}{y_1}\right| < \left|\frac{x_1}{x_2}\right| \tag{23}$$

when the parameter α in (3) *satisfies*

$$0 < \alpha \leq 3. \tag{24}$$

Proof. Applying the inequality $\sqrt{(\beta_1 - \beta_2)^2 + 4} \leq (\beta_1 - \beta_2) + 2$ to (21) yields

$$\frac{y_2}{y_1} \leq \beta_1 - \beta_2 + 1.$$

If $|x_1| = 1$ and $x_2 = 0$, then $|y_2/y_1| < \infty = |x_1/x_2|$ for any choice of $\alpha > 0$. Thus (23) certainly holds when α satisfies (24).

If $x_2 \neq 0$, it follows from (16) that

$$\begin{aligned}\frac{y_2}{y_1} - 1 &\leq \frac{\alpha}{3}(x_1^2 - x_2^2)\\ &= \frac{\alpha}{3}(|x_1| - |x_2|)(|x_1| + |x_2|)\\ &= \frac{\alpha}{3}\Big[|x_2|(|x_1| + |x_2|)\Big]\left(\left|\frac{x_1}{x_2}\right| - 1\right)\\ &\leq \frac{\alpha}{3}\left(\frac{x_1^2 + x_2^2}{2} + x_2^2\right)\left(\left|\frac{x_1}{x_2}\right| - 1\right)\\ &= \frac{\alpha}{6}(1 + 2x_2^2)\left(\left|\frac{x_1}{x_2}\right| - 1\right).\end{aligned}$$

Since $x_1^2 + x_2^2 = 1$ and $|x_1| > |x_2|$, x_2^2 must be less than 1/2. Consequently,

$$\frac{y_2}{y_1} - 1 < \frac{\alpha}{3}\left(\left|\frac{x_1}{x_2}\right| - 1\right).$$

Thus (23) holds if $\alpha \leq 3$. □

The upper bound for α established in Theorem 2 is slightly pessimistic because our experiments show that the SCF iteration converges for α as large as 6.0. However, it is not terribly loose because our experiments also show that convergence failure occurs when $\alpha = 6.5$.

4.2. The more general case.

Our analysis of the SCF iteration for the 2 × 2 problem relies heavily on the symmetry property of the problem and the fact that the solution to the nonlinear eigenvalue problem satisfies $|x_1| = |x_2|$. It is difficult to apply this approach to the more general case in which $n > 2$ and $k > 1$.

Instead of tracking how eigenvectors of $H(X)$ vary from one iteration to another, we will focus in this section on the change in charge density $\rho(X)$. We will use a technique developed in [9] to characterize the mapping between the input charge density used to construct $H(X)$ in (3) and the output charge density obtained directly from the desired eigenvectors of $H(X)$ via (2). We will show that under certain conditions this mapping becomes a contraction when $\alpha < \alpha_{\max}$ for some $\alpha_{\max}$ that depends on the minimum gap between the kth and the $k+1$st eigenvalues of $H(X)$, the norm of L^{-1}, and the problem size n.

We will again assume that there is a gap between the kth and $k+1$st eigenvalues of $H(X)$ for all $X \in \mathbb{R}^{n\times k}$ that satisfies $X^TX = I_k$, and this gap is larger than some lower bound $\delta > 0$. (This is the UWP condition defined in [1].) The significance of this gap will become clear in the following.

Suppose the eigenvalues of $H(X)$ are

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k < \lambda_{k+1} \leq \cdots \leq \lambda_n,$$

for a given X that satisfies $X^TX = I_k$, and the corresponding eigenvectors are $y_1, y_2, \ldots, y_n$. By definition, the density matrix associated with $Y = (y_1, y_2, \ldots, y_k)$ is

$$D(Y) = YY^T.$$

An alternative way to represent this density matrix is

$$D = Z\Omega Z^T,$$

where $Z = (y_1, y_2, \ldots, y_n)$ and $\Omega = \mathrm{Diag}(\underbrace{1, 1, \ldots, 1}_{k}, 0, \ldots, 0)$.

Because $\lambda_k < \lambda_{k+1}$, we can construct a filter function $\phi(\lambda)$ that satisfies

$$\phi(\lambda) = \begin{cases} 1 & \text{for } \ \lambda = \lambda_1, \lambda_2, \ldots, \lambda_k, \\ 0 & \text{for } \ \lambda = \lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_n. \end{cases} \tag{25}$$

If $\phi(\lambda)$ is continuous and differentiable, then we can represent the charge density, which is normally defined as

$$\rho(Y) = \mathrm{diag}(D(Y)),$$

in an alternative form given by

$$\rho = \mathrm{diag}(\phi(H)).$$

If H is constructed from the charge density ρ_{in}, then

$$\rho_{\text{out}} = \mathrm{diag}[\phi(H(\rho_{\text{in}}))]$$

defines a mapping η from ρ_{in} to ρ_{out}, and this is the mapping implicitly constructed at each SCF iteration.

We would like to identify the condition under which η becomes a contraction. Such a condition will ensure that the SCF iteration converges to a fixed point of η that is the solution to our nonlinear eigenvalue problem.

To seek such a condition, we will show that

$$\|\eta(\rho_1) - \eta(\rho_2)\|_1 < \gamma \|\rho_1 - \rho_2\|_1 \tag{26}$$

for any ρ_1 and ρ_2 that satisfy the standard definition (2), and identify the requirement under which $\gamma < 1$.

Constructing a proper filter function is the key to proving (26). We will choose $\phi(t)$ to be a Fermi–Dirac distribution [4] of the form

$$\phi(t) = f_\mu(t) = \frac{1}{1 + e^{\beta(t-\mu)}}, \tag{27}$$

where μ is implicitly determined by the input matrix argument to $\phi(t)$ and $\beta > 0$ is a constant. To be specific, μ is the solution of the equation

$$\mathrm{trace}(\phi(H)) = \mathrm{trace}(f_\mu(H)) = k. \tag{28}$$

Because $\sum_{i=1}^{n} f_\mu(\lambda_i)$ is monotonic with respect to μ for a fixed β, the solution to (28) is unique for any choice of β and H. Figure 5 shows how Fermi–Dirac distributions look with different β values and $\mu = 0$. Notice that a larger β value leads to a sharper drop-off of $\phi(t)$ from 1 to 0.

If the UWP condition holds, then there exists a constant β sufficiently large so that (25) is fulfilled in finite precision arithmetic.

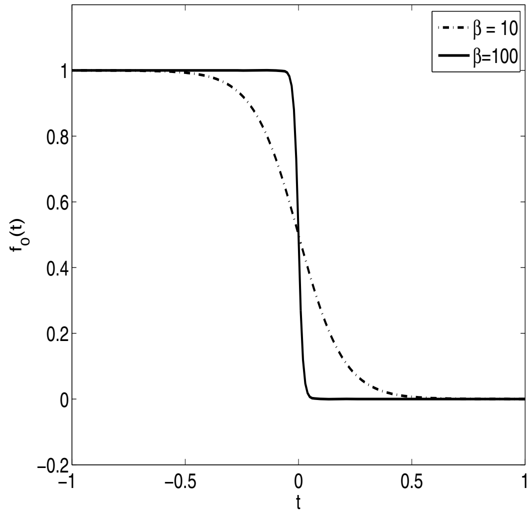


FIG. 5. *Fermi–Dirac distribution* $f_\mu(t) = \frac{1}{1+e^{\beta(t-\mu)}}$ *for* $\mu = 0$.

Let H_1 and H_2 be Hamiltonians constructed from the charge densities ρ_1 and ρ_2, respectively. It is easy to see that

$$\begin{aligned}\|\eta(\rho_1) - \eta(\rho_2)\|_1 &= \|\mathrm{diag}[f_{\mu_1}(H_1)] - \mathrm{diag}[f_{\mu_2}(H_2)]\|_1 \\ &\le \|\mathrm{diag}[f_{\mu_1}(H_1) - f_{\mu_2}(H_1)]\|_1 + \|\mathrm{diag}[f_{\mu_2}(H_1) - f_{\mu_2}(H_2)]\|_1. \end{aligned} \tag{29}$$

Without loss of generality, let us assume $\mu_1 \ge \mu_2$. As a result, $f_{\mu_1}(t) \ge f_{\mu_2}(t)$ for any t. Hence

$$\begin{aligned}\|\mathrm{diag}[f_{\mu_1}(H_1) - f_{\mu_2}(H_1)]\|_1 &= \mathrm{trace}[f_{\mu_1}(H_1) - f_{\mu_2}(H_1)] \\ &= \mathrm{trace}[f_{\mu_1}(H_1)] - \mathrm{trace}[f_{\mu_2}(H_1)]. \end{aligned} \tag{30}$$

Since $\mathrm{trace}[f_{\mu_1}(H_1)] = \mathrm{trace}[f_{\mu_2}(H_2)] = k$, it is easy to see that

$$\begin{aligned}\mathrm{trace}[f_{\mu_1}(H_1)] - \mathrm{trace}[f_{\mu_2}(H_1)] &= \mathrm{trace}[f_{\mu_2}(H_2)] - \mathrm{trace}[f_{\mu_2}(H_1)] \\ &= \mathrm{trace}[f_{\mu_2}(H_2) - f_{\mu_2}(H_1)] \\ &\le \|\mathrm{diag}[f_{\mu_2}(H_2) - f_{\mu_2}(H_1)]\|_1. \end{aligned} \tag{31}$$

Consequently, it follows from (29), (30), and (31) that

$$\begin{aligned}\|\eta(\rho_1) - \eta(\rho_2)\|_1 &\le 2\|\mathrm{diag}[f_{\mu_2}(H_2) - f_{\mu_2}(H_1)]\|_1 \\ &\le 2n\|f_{\mu_2}(H_2) - f_{\mu_2}(H_1)\|_1. \end{aligned} \tag{32}$$

Now to show (26) and to derive an upper bound for α, all we need to do is show that

$$\|f_{\mu_2}(H_2) - f_{\mu_2}(H_1)\|_1 < \frac{\gamma}{2n}\|\rho_1 - \rho_2\|_1$$

for some γ that is proportional to α. Before we do that, we will first prove the following lemma, which allows us to establish a desirable relationship between $f_{\mu_2}(H_2) - f_{\mu_2}(H_1)$ and $H_2 - H_1$.

LEMMA 4. *Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices, and let $f(t)$ be the Fermi–Dirac distribution defined in* (27). *Suppose $A = V_A D_A V_A^T$ and $B = V_B D_B V_B^T$ are the spectral decompositions of A and B, respectively, i.e., $V_A^T V_A = V_B^T V_B = I$ and*

$$D_A = \begin{pmatrix} \lambda_1^A & & & \\ & \lambda_2^A & & \\ & & \ddots & \\ & & & \lambda_n^A \end{pmatrix}, \quad D_B = \begin{pmatrix} \lambda_1^B & & & \\ & \lambda_2^B & & \\ & & \ddots & \\ & & & \lambda_n^B \end{pmatrix}.$$

Then the identity

$$f(A) - f(B) = V_A(C \odot \Delta)V_B^T$$

holds, where $\Delta = V_A^T(A - B)V_B$, the (j,k)th entry of the matrix C is defined by

$$C_{j,k} = \begin{cases} \frac{f(\lambda_j^A)-f(\lambda_k^B)}{\lambda_j^A-\lambda_k^B} & \text{if } \lambda_j^A \neq \lambda_k^B, \\ f'(\lambda) & \text{if } \lambda_j^A = \lambda_k^B = \lambda, \end{cases}$$

and $C \odot \Delta$ denotes the Hadamard product of C and Δ.

Proof. It follows from the matrix version of the Cauchy integral formula [3] that

$$f(A) - f(B) = \frac{1}{2\pi i}\oint_\Gamma f(z)\left[(zI - A)^{-1} - (zI - B)^{-1}\right]dz, \tag{33}$$

where Γ is a closed contour that contains the spectra of both A and B.

Using the identity

$$(zI - A)^{-1} - (zI - B)^{-1} = (zI - A)^{-1}(A - B)(zI - B)^{-1},$$

we can express the right-hand side of (33) as

$$\begin{aligned} &\frac{1}{2\pi i}\oint_\Gamma f(z)V_A(zI - D_A)^{-1}V_A^T(A - B)V_B(zI - D_B)^{-1}V_B^T dz \\ &\quad = \frac{1}{2\pi i}\oint_\Gamma f(z)V_A[(w_A(z)w_B(z)^T) \odot \Delta]V_B^T dz, \end{aligned} \tag{34}$$

where $w_A = \text{diag}[(zI - D_A)^{-1}]$, $w_B = \text{diag}[(zI - D_B)^{-1}]$.

Since the only term in (34) that contains z is $w_A(z)w_B(z)^T$, it follows that

$$f(A) - f(B) = V_A\left[\left(\frac{1}{2\pi i}\oint_\Gamma f(z)w_A(z)w_B(z)^T dz\right) \odot \Delta\right]V_B^T.$$

Let

$$C = \frac{1}{2\pi i}\oint_\Gamma f(z)w_A(z)w_B(z)^T dz.$$

It is easy to verify that the (j,k)th entry of C can be expressed as

$$C_{j,k} = \frac{1}{2\pi i}\oint_\Gamma \frac{f(z)}{(z - \lambda_j^A)(z - \lambda_k^B)}dz. \tag{35}$$

If $\lambda_j^A \neq \lambda_k^B$, the expression above can be evaluated as

$$C_{j,k} = \frac{1}{2\pi i}\frac{1}{\lambda_j^A - \lambda_k^B}\oint_\Gamma \left(\frac{f(z)}{z-\lambda_j^A} - \frac{f(z)}{z-\lambda_k^B}\right)dz. \tag{36}$$

If $\lambda_j^A = \lambda_k^B = \lambda$, (35) becomes

$$C_{j,k} = \frac{1}{2\pi i}\oint_\Gamma \frac{f(z)}{(z-\lambda)^2}dz. \tag{37}$$

Invoking the scalar version of the Cauchy integral formula in both (36) and (37), we then obtain

$$C_{j,k} = \begin{cases} \frac{f(\lambda_j^A)-f(\lambda_k^B)}{\lambda_j^A-\lambda_k^B} & \text{if } \lambda_j^A \neq \lambda_k^B, \\ f'(\lambda) & \text{if } \lambda_j^A = \lambda_k^B = \lambda. \end{cases} \qquad \square$$

Suppose $H_1 = X_1\Lambda_1X_1^T$ and $H_2 = X_2\Lambda_2X_2^T$ are the spectral decompositions of H_1 and H_2, respectively. A direct application of Lemma 4 to H_1 and H_2 yields

$$\begin{aligned} \|f_{\mu_2}(H_2) - f_{\mu_2}(H_1)\|_1 &= \|X_2[C\odot(X_2^T(H_2-H_1)X_1)]X_1^T\|_1 \\ &\leq n\|C\odot(X_2^T(H_2-H_1)X_1)\|_1 \\ &\leq n^2\|C\|_1\|H_2-H_1\|_1 \\ &\leq \alpha n^2\|C\|_1\|L^{-1}\|_1\|\rho_2-\rho_1\|_1. \end{aligned} \tag{38}$$

To establish an upper bound for $\|C\|_1$, we can use the mean value theorem and the fact that

$$|f_\mu'(t)| = \left|\frac{-\beta e^{\beta(t-\mu)}}{(1+e^{\beta(t-\mu)})^2}\right| \leq \frac{\beta}{4}$$

to first show that

$$\max_{j,k}|C_{j,k}| \leq \beta/4.$$

It follows immediately that

$$\|C\|_1 \leq n\beta/4. \tag{39}$$

Combining (32), (38), and (39), we obtain

$$\|\eta(\rho_2)-\eta(\rho_1)\|_1 \leq \frac{\alpha n^4\beta\|L^{-1}\|_1}{2}\|\rho_2-\rho_1\|_1.$$

We can easily see that η is a contraction if α satisfies

$$\alpha < \frac{2}{n^4\beta\|L^{-1}\|_1}. \tag{40}$$

It may seem surprising that the upper bound that ensures $\eta(\rho)$ becomes a contraction depends on a parameter β that is present in neither the original eigenvalue problem (1) nor the description of the SCF iteration. However, if we go back to Figure 5 and recall that the choice of β is dictated by the smallest gap between $\lambda_k(H)$ and

$\lambda_{k+1}(H)$ for all valid H matrices, then it becomes clear that the dependency of (40) on β simply says that for problems in which the gap between $\lambda_k(H)$ and $\lambda_{k+1}(H)$ is small, a smaller upper bound of α is required to ensure that the SCF iteration converges from any starting point.

We should point out that the bound established in (40) is pessimistic. In particular, the n^4 factor on the denominator, which is introduced by the use of a loose inequality in (32) and the use of 1-norms to bound the norms of the orthogonal matrices X_1 and X_2 in (38), is rather conservative. In our numerical experiments, we observed that the SCF iteration may converge for α values that are much larger than the right-hand side of (40). However, the qualitative behavior of the SCF iteration seems to be correctly characterized by (40). Table 1 shows both the experimentally observed largest α values (α_1) for which the SCF iteration converges and the experimentally observed smallest α values (α_2) for which the SCF iteration fails to converge for problems with different choices of n and k. The optimal bound lies within the interval (α_1, α_2). We can clearly see that the optimal bound for α decreases as n increases. For the same value of n, changing the value of k in Table 1 results in a change of the gap $\lambda_{k+1} - \lambda_k$. For each combination of n and k, the smallest gap among the various choices of α's that we experimented with is shown in Table 1. The last two rows of Table 1 clearly indicate that for the same n, a smaller $\lambda_{k+1} - \lambda_k$, which corresponds to a larger β value in (40), leads to a more restrictive choice of α for which the SCF iteration is guaranteed to converge.

TABLE 1

Observation from numerical experiments performed to determine the optimal bound for α. In these experiments, the L matrix in (3) *is constructed as the one-dimensional discrete Laplacian with* 2 *on the diagonal and* -1 *on the sub- and sup-diagonals. The dimension of the matrix is n. We look for k smallest eigenvalues and the corresponding eigenvectors. The SCF iteration converges for* $\alpha \leq \alpha_1$ *and fails to converge for* $\alpha \geq \alpha_2$. *This implies that the optimal bound for* α *lies in* (α_1, α_2). *The spectral gap* $\lambda_{k+1} - \lambda_k$ *listed here is smallest among all gaps associated with the different choices of* α *values that we experimented with. These gaps were computed using a trust-region enabled SCF iteration discussed in* [11].

n	k	$\lambda_{k+1} - \lambda_k$	$\|L^{-1}\|_1$	α_1	α_2
2	1	2.0	1.0	6.0	6.5
10	2	0.37	15.0	0.8	0.9
100	10	0.02	1275.0	0.05	0.06
100	4	0.0087	1275.0	0.002	0.0025

In general, the minimum gap between $\lambda_k(H)$ and $\lambda_{k+1}(H)$ is not known a priori. However, when α is sufficiently small, we can estimate such a gap by calculating the difference between the kth and $k+1$st eigenvalues of L. Such an estimate can in turn be used to derive a suitable β value that would allow (27) to achieve the filtering effect (25) in finite precision arithmetic.

5. Concluding remarks.

We examined the convergence of the self-consistent field (SCF) iteration used to solve a class of nonlinear eigenvalue problems defined in (1). Our analysis shows that for this type of problem the SCF iteration produces a sequence of approximate solutions $X^{(i)}$ that contain two convergent subsequences. However, the limit points associated with these convergent subsequences may be different, as we demonstrated in a numerical example. We identified the condition under which the SCF iteration becomes a contractive fixed point iteration that will converge to the solution of the nonlinear eigenvalue problem. Our main result suggests that this condition can be characterized by an upper bound placed on the parameter α in (1).

In the most general case, the upper bound we derived characterizes the qualitative behavior of the SCF iteration, although the bound itself is somewhat pessimistic. When the dimension of the problem is 2×2, we can give a much tighter bound using a completely different technique. To generalize such a bound for the Hartree–Fock (HF) or the Kohn–Sham (KS) problem, we need to analyze the relative contribution of the exchange and exchange-correlation terms to the HF and KS Hamiltonians, respectively. We will pursue such analysis in future research.

Acknowledgments. We would like to thank the anonymous referees and the associate editor for careful reading and helpful comments.

REFERENCES

[1] C. Le Bris, *Computational chemistry from the perspective of numerical analysis*, Acta Numer., 14 (2005), pp. 363–444.

[2] E. Cancès and C. Le Bris, *On the convergence of SCF algorithms for the Hartree-Fock equations*, Math. Model. Numer. Anal., 34 (2000), pp. 749–774.

[3] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[4] C. Kittel and H. Kroemer, *Thermal Physics*, W. H. Freeman, San Francisco, 1980.

[5] J. Koutecký and V. Bonacic, *On convergence difficulties in the iterative Hartree-Fock procedure*, J. Chem. Phys., 55 (1971), pp. 2408–2413.

[6] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, Cambridge, UK, 2004.

[7] J. Nocedal and S. Wright, *Numerical Optimization*, Springer-Verlag, New York, 1999.

[8] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[9] E. Prodan and P. Nordlander, *On the Kohn–Sham equations with periodic background potentials*, J. Statist. Phys., 111 (2003), pp. 967–992.

[10] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: An Introduction to Advanced Electronic Structure Theory*, Dover, New York, 1996.

[11] C. Yang, J. C. Meza, and L.-W. Wang, *A trust region direct constrained minimization algorithm for the Kohn–Sham equation*, SIAM J. Sci. Comput., 29 (2007), pp. 1854–1875.